# AGAIN: Automatic desiGn for totAl chemIcal synthesis of proteiNs

Yufeng Xie

Tsinghua University

December 28, 2019

# GAIN


Figure: Designing a mutant Dpo4 for total chemical synthesis(GAIN)[1].

[1] Xu et al. 2017.

# Benchmark

Dpo4-1: $His_6 - Nle^{1} - Gly^{41} - NHNH_2$
Dpo4-2: $Cys^{42} - Val^{43} - Ser^{85} - NHNH_2$
Dpo4-3: $Acm - Cys^{86} - Arg^{87} - Tyr^{122} - NHNH_2$
Dpo4-4: $Acm - Cys^{123} - Leu^{124} - Ala^{154} - NHNH_2$
Dpo4-5: $Tfa - Thz - Asp^{156} - Leu^{206} - NHNH_2$
Dpo4-6: $Cys^{207} - Ile^{208} - Leu^{228} - NHNH_2$
Dpo4-7: $Tfa - Thz - Asp^{230} - Arg^{267} - NHNH_2$
Dpo4-8: $Tfa - Thz - Ile^{269} - Tyr^{312} - NHNH_2$
Dpo4-9: $Cys^{313} - Glu^{314} - Thr^{352} - NHNH_2$[2]

[2] Xu et al. 2017.

# Considerations

Aim:

- Feasible: chemical synthesis
  - native chemical ligation site: Cys, Ala
  - more ligation site: point mutations
- Economical: Ile

mutiple sequence alignment(MSA) $\rightarrow$ mutate to be Ala $\rightarrow$ double check: active site, bond distinguish

Two views:

- MSA-based method: conservative, but successful $\leftarrow$ protein evolution
- MSA-independent method: simulated evolution $\rightarrow$ find more sites

# MSA-based method

mutiple sequence alignment(MSA):

- BLAST(server)
- ClustalW[3], hmmer[4](localhost)
  - ClustalW(profile)
  - hmmer(hits)

```
1       CLUSTAL 2.1 multiple sequence alignment
2
3
4       d1dwta_         -GLSDGEWQQVLNVWGKVEA--DIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKA
5       d1emy__         -GLSDGEWELVLKTWGKVEA--DIPGHGETVFVRLFTGHPETLEKFDKFKHLKTEGEMKA
6       d1mwca_         -GLSDGEWQLVLNVWGKVEA--DVAGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKA
7       d2mm1__         -GLSDGEWQLVLNVWGKVEA--DIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKA
8       d1a6m__         -VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKA
9       d1mbs__         -GLSDGEWHLVLNVWGKVET--DLAGHGQEVLIRLFKSHPETLEKFDKFKHLKSEDDMRR
10      d1lht__         -GLSDDEWNHVLGIWAKVEP--DLSAHGQEVIIRLFQLHPETQERFAKFKNLTTIDALKS
11      d1myt__         -----ADFDAVLKCWGPVEA--DYTTMGGLVLTRLFKEHPETQKLFPKFAGIAQAD-IAG
12      d1vrea_         -GLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHDMAAVFG-FSGASDPGVADL
13      d2hbg__         -GLSAAQRQVIAATWKDIAGADNGAGVGKKCLIKFLSAHPQMAAVFG-FSGASDPGVAAL
14      d1eco__         --LSADQISTVQASFDKVKG--DPVG----ILYAVFKADPSIMAKFTQFAGK-DLESIKG
15      d1mba__         -SLSAAEADLAGKSWAPVFA--NKNANGLDFLVALFEKFPDSANFFADFKGK-SVADIKA
16      d2gdm__         GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNP
17                            :       :  .    :        :  .:         *  :
```

Figure: MSA of 13 globins sequences

[3] Sievers et al. 2011.

[4] Eddy 1998.

# MSA-based method

double check: pdb file → scripts → distance map

```
REMARK coordinates from energy minimization
REMARK rmsd bonds= 0.004389  rmsd angles=  1.50667
REMARK nonbonded cutoff= 13 Angstroms  dieletric= 1
REMARK cycles= 5 steps= 200
REMARK parameter file 1  : CNS_TOPPAR:protein.param
REMARK molecular structure file: psipred_6_dgsa_3A_h2_18_7_h.mtf
REMARK input coordinates: psipred_6_dgsa_3A_h2_18_7_h.pdb
REMARK ncs= none
REMARK FILENAME="psipred_6_dgsa_3A_h2_18_7_hMIN.pdb"
REMARK DATE:11-Mar-2019  17:05:26       created by user: kpb3
REMARK VERSION:1.21
ATOM      1  CB  THR   228     -11.225   4.038   4.307  1.00  0.00
ATOM      2  OG1 THR   228     -11.900   3.553   5.475  1.00  0.00
ATOM      3  HG1 THR   228     -11.752   4.220   6.164  1.00  0.00
ATOM      4  CG2 THR   228     -11.812   5.384   3.929  1.00  0.00
ATOM      5  C   THR   228     -10.589   1.778   3.452  1.00  0.00
ATOM      6  O   THR   228     -10.963   0.976   4.311  1.00  0.00
ATOM      7  HT1 THR   228     -12.996   1.877   2.354  1.00  0.00
ATOM      8  HT2 THR   228     -13.401   3.500   2.825  1.00  0.00
ATOM      9  N   THR   228     -12.838   2.655   3.025  1.00  0.00
ATOM     10  HT3 THR   228     -13.027   2.323   3.995  1.00  0.00
ATOM     11  CA  THR   228     -11.408   3.031   3.159  1.00  0.00
ATOM     12  N   TYR   229      -9.503   1.580   2.714  1.00  0.00
ATOM     13  H   TYR   229      -9.222   2.256   2.049  1.00  0.00
ATOM     14  CA  TYR   229      -8.659   0.410   2.927  1.00  0.00
ATOM     15  CB  TYR   229      -8.509  -0.409   1.640  1.00  0.00
ATOM     16  CG  TYR   229      -8.234   0.383   0.380  1.00  0.00
```

Figure: A pdb demo file

# MSA-based method

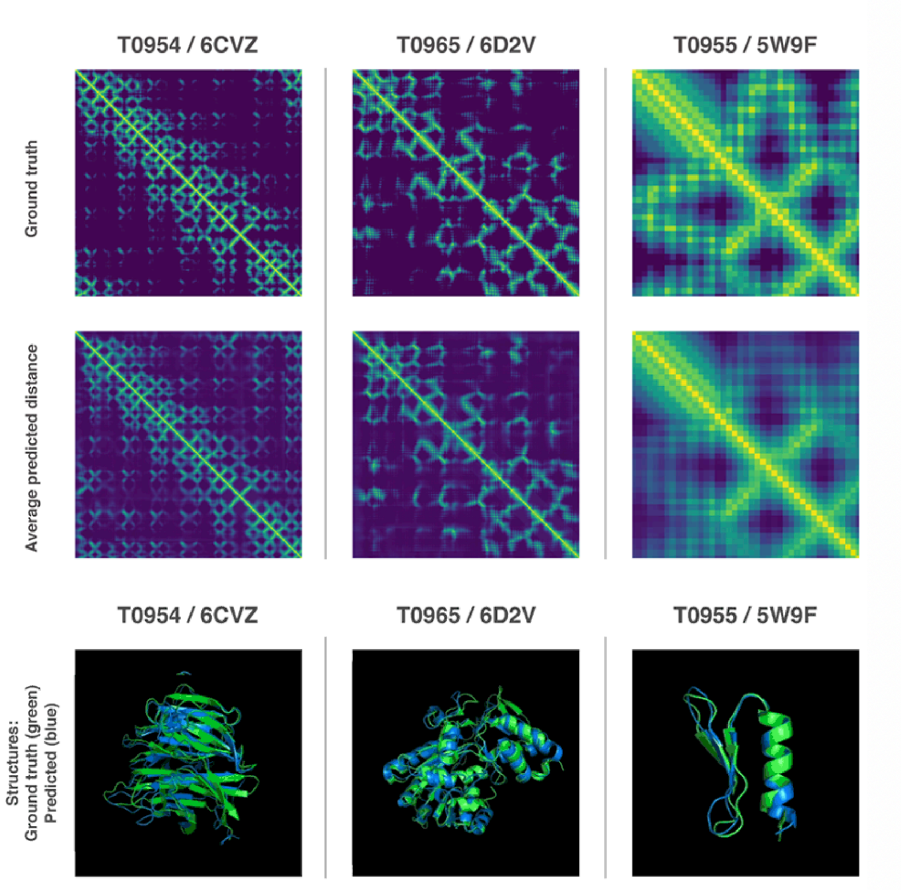



Figure: distance map[5]

to distinguish hydrogen bond..

[5] https://deepmind.com/blog/article/alphafold

# Why MSA-independent method

- to overcome site limitation problem
  - short evolution history
  - speices-specific
  - small protein family
  - sequence uncovered
- problem exist? $\rightarrow$ theoretically possible $\rightarrow$ alternative method
- interest

# Distance map

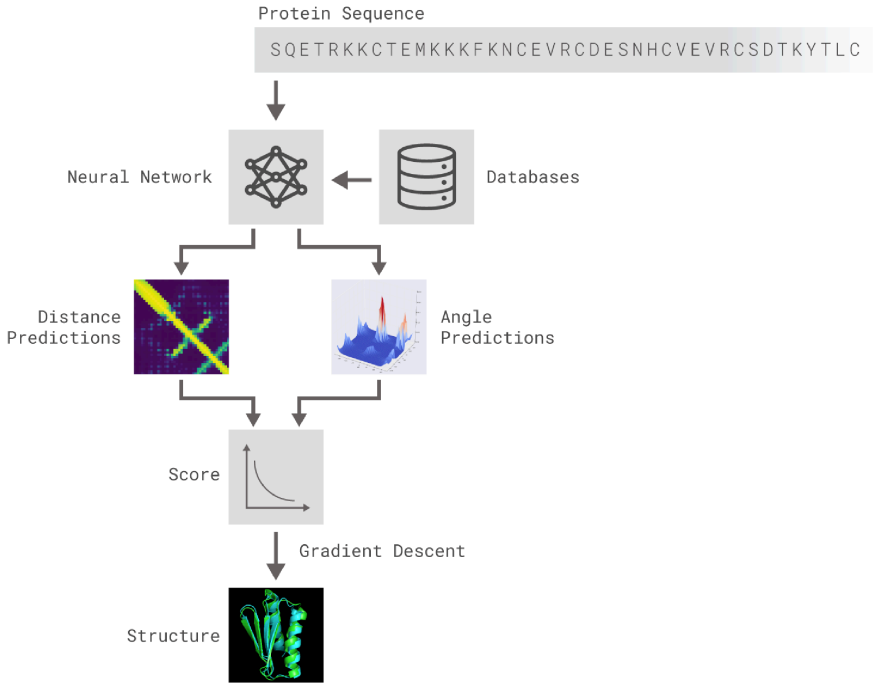



Figure: Alphafold[6]

[6] https://deepmind.com/blog/article/alphafold

# Distance map

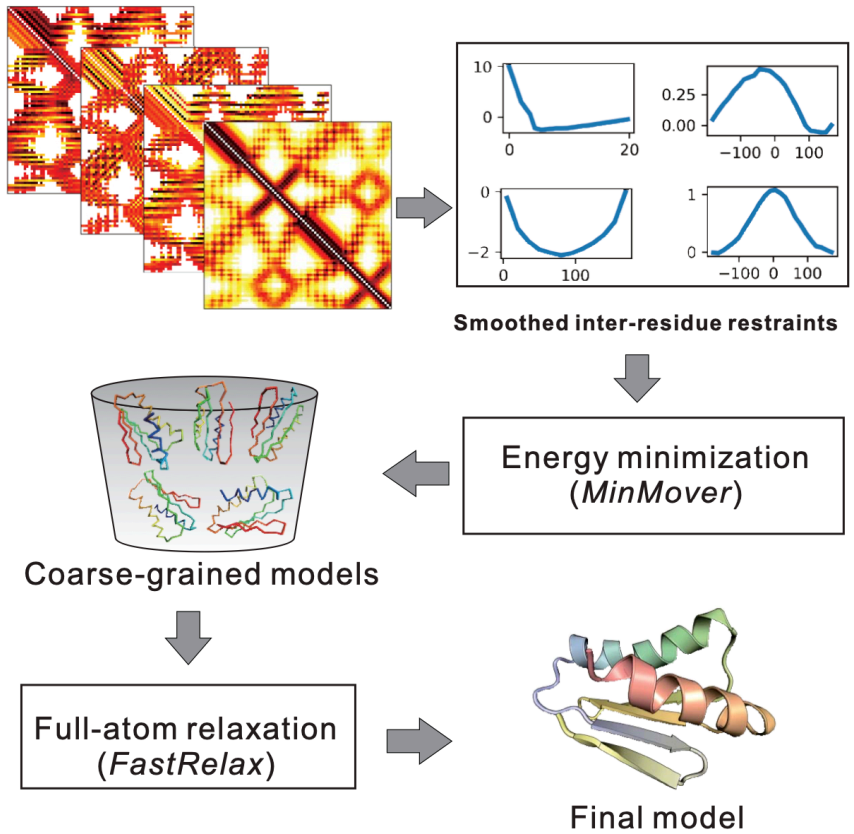
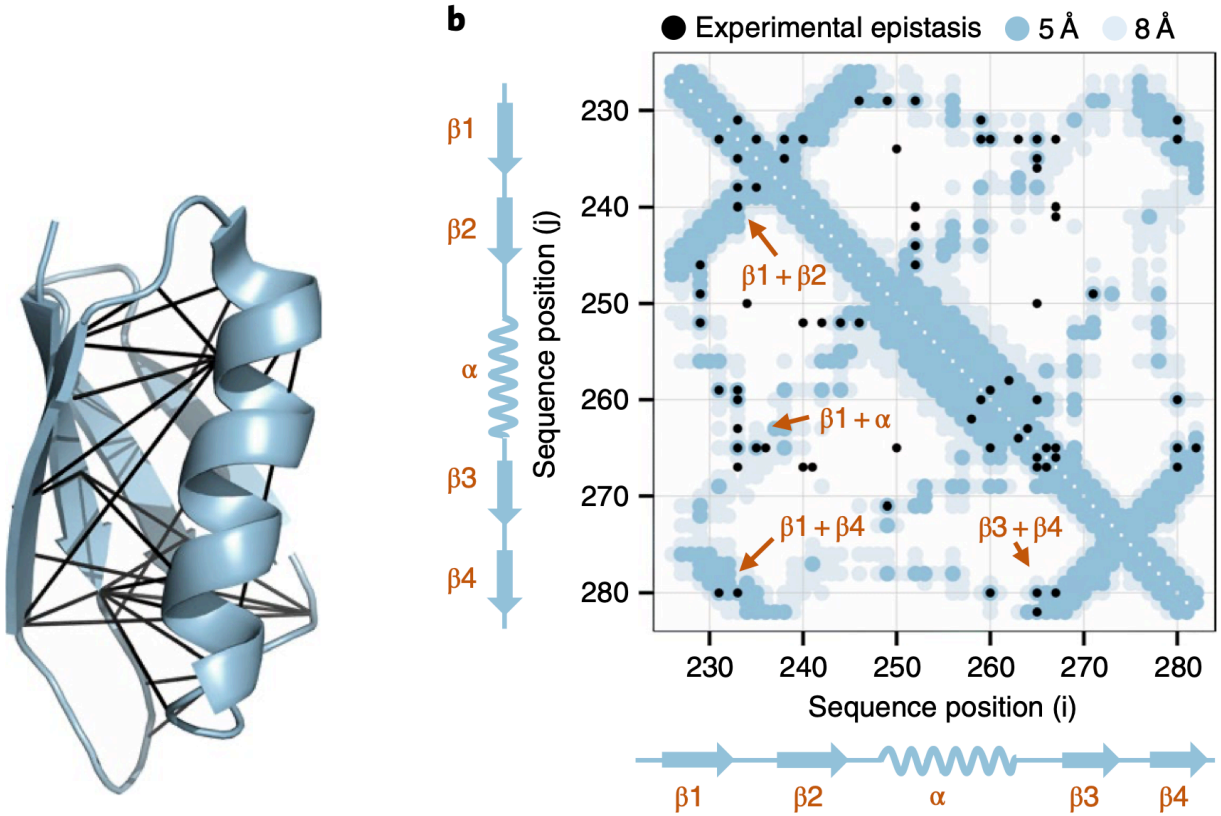



Figure: trRosetta[7]

[7] Yang et al. 2019.

# Distance map is powerful

what is the relationship between maps and mutations?




(a) Mutations (b) Distance map

Figure: Infer distance map from mutations[8]

[8] Rollins et al. 2019.

# Relationship

mutations $\rightarrow$ maps
maps $\rightarrow$ mutations?
more reading and thinking...

THANKS!